MOSSTrack : Modality-Specific Spatio-Temporal Context Learning for RGB-T Tracking

Supplementary Material

In this supplementary material, we provide additional content to complement the main manuscript:

- More implementation details.
- Introduction of benchmarks
- More comparisons with state-of-the-art methods

1. More implementation details

This section provides implementation details of the model that were omitted from the main manuscript due to space limitations.

1.1. Model

We adopt HiViT-Base [24] as the encoder of our model. To mitigate the loss of spatial information caused by a large one-step downsampling, we employ a progressive patch embedding strategy. Unlike the vanilla ViT [3], which directly applies a 16×16 patch embedding, the template images $Z_r, Z_t \in \mathbb{R}^{3\times H_z\times W_z}$ and search images $X_r, X_t \in \mathbb{R}^{3\times H_x\times W_x}$ from both modalities are processed through three downsampling stages: one 4×4 patch embedding layer followed by two 2×2 merging layers. This hierarchical design yields final patch embeddings with an effective 16×16 resolution while preserving substantially richer spatial details. As a result, we obtain the template tokens $H_r^z, H_t^z \in \mathbb{R}^{N_z\times D}$ and the search tokens $H_r^x, H_t^x \in \mathbb{R}^{N_x\times D}$, where $N_z = H_z W_z/16^2$ and $N_x = H_x W_x/16^2$. The feature dimension of all tokens is set to $D = 512$.

Table 1. Comparison of tracking performance, model parameters, and computational cost.

Model	RGBT234		LasHeR			Params	FLOPs
	PR(%)	SR(%)	PR(%)	NPR(%)	SR(%)		
SDSTrack [5]	84.8	62.5	66.5	62.7	53.1	102M	108G
STTrack [6]	89.8	66.7	76.0	71.9	60.3	128M	91G
MOSSTrack	94.3	71.5	76.9	72.6	61.2	78M	65G

In addition, we compare MOSSTrack with two recently proposed high-performing RGB-T trackers, SDSTrack [5] and STTrack [6], in terms of tracking performance (PR/SR scores), parameter count (Params), and computational cost (FLOPs). As shown in Table 1, our method achieves clearly superior tracking accuracy while using fewer parameters and requiring lower computational overhead. This indicates that MOSSTrack attains a more favorable balance between performance and efficiency compared with existing advanced approaches.

1.2. Prediction Head

The prediction head of the proposed MOSSTrack follows the design of OSTrack [22], with a center-based head employed to estimate the target center and scale. Specifically, the output search-region tokens are fed into three fully convolutional networks to generate the target classification score map $S \in [0,1]^{1\times\frac{H_x}{P}\times\frac{W_x}{P}}$, the local offset map $O \in [0,1)^{2\times\frac{H_x}{P}\times\frac{W_x}{P}}$ and the bounding box size map $B \in [0,1]^{2\times\frac{H_x}{P}\times\frac{W_x}{P}}$. The location with the highest score in S is selected as the target center, and the final tracking result is obtained by combining its corresponding offset O and bounding box size B.

$$(x_i, y_i) = \text{argmax}_{(x,y)} S_{xy}, \tag{1}$$

$$(x, y) = (x_i + O(0, x_i, y_i), y_d + O(1, x_i, y_i)), \tag{2}$$

$$(w, h) = (B(0, x_i, y_i), B(1, x_i, y_i)). \tag{3}$$

Here, x, y, w and h denote the predicted bounding box center coordinates and its width and height, respectively.

1.3. Training and Inference

During training, we apply several standard data augmentation strategies, including horizontal flipping, spatial jittering, and brightness/color variations, to improve the model's generalization ability. The Transformer encoder is pretrained on large-scale RGB tracking datasets—LaSOT [4], GOT-10K [7], COCO [14], and TrackingNet [18]—to endow the model with strong general-purpose visual feature representations. Building upon this pretrained backbone, we integrate all proposed components and further train the model on the training subset of LasHeR [13], allowing the tracker to learn cross-modal interactions and modality-specific cues essential for RGB-T tracking. Throughout training, focal loss [15] is employed for classification to mitigate class imbalance, while the regression branch adopts a combination of generalized IoU loss and ℓ_1 loss [19] to enhance both bounding-box overlap quality and localization accuracy. The overall loss function is formulated as follows:

$$L = L_{cls} + \lambda_{iou} L_{iou} + \lambda_{L_1} L_1, \tag{4}$$

where λ_{iou} and λ_{L_1} are regularization parameters, which are set to 2 and 5, respectively, in our experiments.

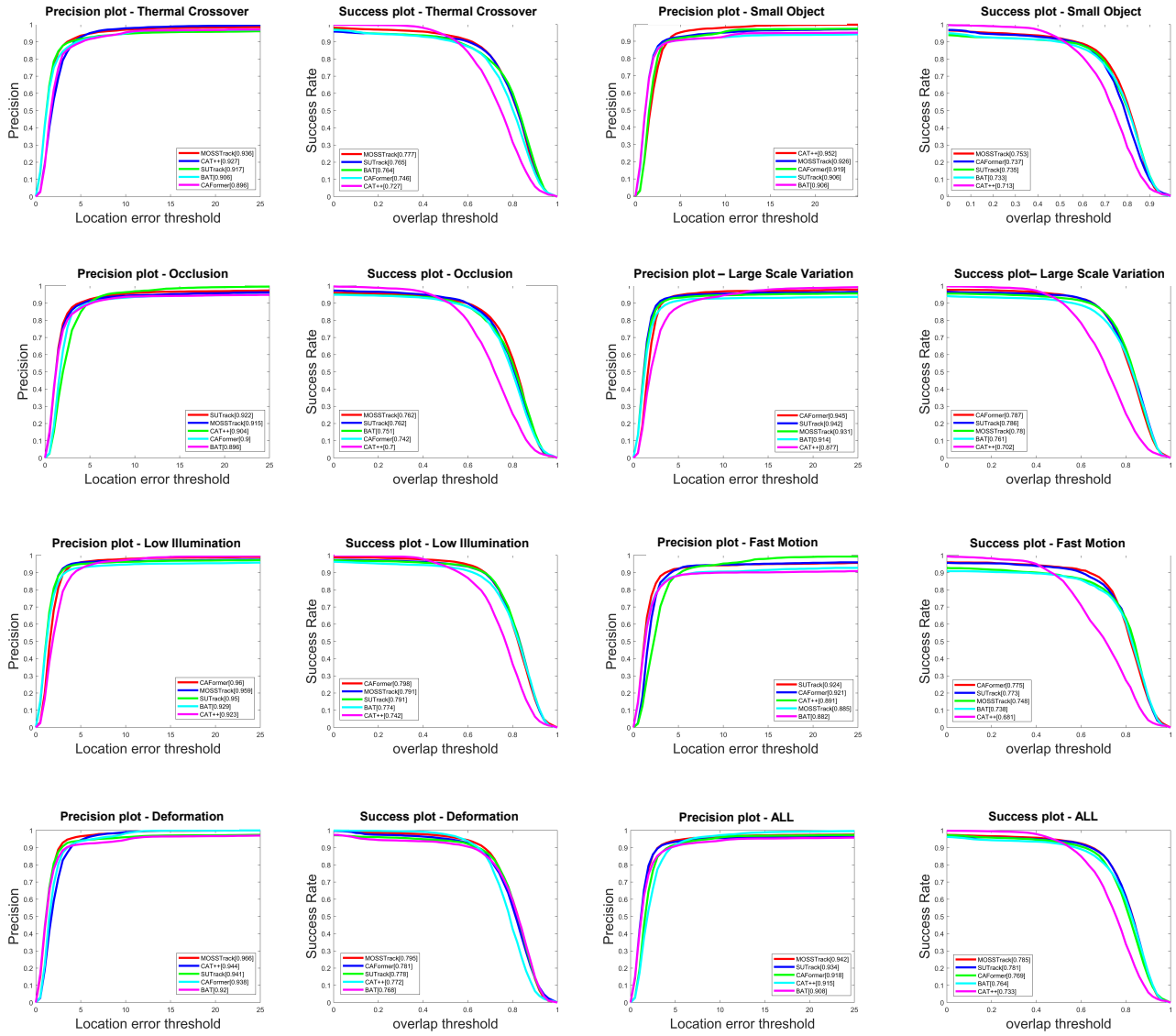

Figure 1. Comparison of the proposed MOSSTrack with other advanced trackers on the GTOT dataset across seven challenging attribute subsets. Overall success and precision plots on the full GTOT test set are also presented.

During inference, the temporal tokens are updated on a per-frame basis, enabling the model to capture subtle variations of the target at each time step. Meanwhile, we retain the conventional template update strategy to establish long-term spatio-temporal associations, preventing error accumulation caused by temporary occlusion or disappearance of the target. The dynamic template is updated only when two conditions are satisfied: the confidence score exceeds 0.7, and the interval between the two frames is greater than 50 frames. Additionally, following standard practice, a Hanning window penalty is applied to the classification score map during inference to incorporate positional priors and reduce jitter and drift in tracking.

2. Introduction of Benchmarks

In this section, we provide a detailed introduction to the datasets used in our paper and the tracking attribute challenges defined for them. In addition, we present GTOT, a small RGB-T tracking dataset used for evaluation.

GTOT. The GTOT [10] dataset contains 50 pairs of RGB–T video sequences, each annotated with target bounding boxes to ensure consistent labeling across modalities. The visible and thermal infrared images are aligned by manually selecting keypoint correspondences and estimating a homography matrix. As one of the earliest RGB–T tracking datasets featuring diverse challenging scenarios, GTOT groups its 50 sequences into seven attribute categories: Oc-

clusion (OCC), Deformation (DEF), Low Illumination (LI), Small Objects (SO), Large Scale Variation (LSV), Thermal Crossover (TC), and Fast Motion (FM). In this supplementary material, we provide an attribute-based performance comparison of MOSSTrack against existing state-of-the-art RGB–T trackers on the GTOT dataset.

RGBT210. The RGBT210 dataset [11] consists of 210 pairs of RGB–T tracking sequences, with some sequences reaching up to 8,000 frames. It provides frame-level occlusion annotations, allowing comprehensive evaluation of a tracker's robustness under different occlusion conditions. Extending the challenge definitions used in GTOT, RGBT210 includes 12 attribute categories and introduces four additional challenges—Motion Blur (MB), Camera Moving (CM), Low Resolution (LR), and Background Clutter (BC). The SO attribute is removed, and the OCC attribute is further divided into three more fine-grained categories: No Occlusion (NO), Partial Occlusion (PO), and Total Occlusion (TO), enabling more detailed attribute-sensitive performance analysis.

RGBT234. The RGBT234 dataset [12] is an extension of RGBT210, comprising 234 paired RGB–T video sequences with approximately 116.6k aligned frame pairs. It includes 22 categories of tracking targets. Similar to RGBT210, the dataset organizes all sequences into 12 commonly used challenge attributes, defined according to environmental conditions and the states of the tracked objects.

Table 2. Tracking challenge attributes present in existing RGB-T datasets.

Attribute	Definition
NO	No Occlusion - the target is not occluded.
PO	Partial Occlusion - the target object is partially occluded.
TO	Total Occlusion - the target object is totally occluded.
HO	Hyaline Occlusion - the target is occluded by hyaline object.
OV	Out-of-View - the target leaves the camera field of view.
LI	Low Illumination - the illumination in the target region is low.
HI	High Illumination - the illumination in the target is too strong to identify the target.
AIV	Abrupt Illumination Variation - the illumination of the target changes significantly.
LR	Low Resolution - the resolution in the target region is low.
DEF	Deformation - non-rigid object deformation.
BC	Background Clutter - the background information which includes the target object is messy.
SA	Similar Appearance - there are objects of similar appearance near the target.
TC	Thermal Crossover - the target has similar temperature with other objects or background surroundings.
MB	Motion Blur - Blurred target due to target movement or camera shake.
CM	Camera Moving - the target object is captured by moving camera.
FL	Frame Lost - some of thermal frames are lost.
FM	Fast Motion - the motion of the ground truth between two adjacent frames is larger than 20 pixels.
SV	Scale Variation - the ratio of the first bounding box and the current bounding box is out of the range [0.5,2].
SO	Small Object - the number of pixels in the ground truth bounding box is less than 400.
TVS	Thermal-Visible Separation - the target positions do not overlap in the visible and thermal images.
ARC	Aspect Ratio Change - the ratio of bounding box aspect is outside the range [0.5,2].

LasHeR. LasHeR [13] is currently one of the largest RGB-T tracking datasets, containing 1,224 RGB-T video pairs, of which 245 sequences constitute the test set and 979 sequences are designated for training. The dataset includes 32 object categories and 19 challenge attributes, comprising over 734.8k aligned RGB-T frame pairs. In addition to the attributes defined in RGBT210 and RGBT234, LasHeR introduces seven new challenges: Hyaline Occlusion (HO), Frame Lost (FL), Abrupt Illumination Variation (AIV), Aspect Ratio Change (ARC), Out-of-View (OV), High Illumination (HI), and Similar Appearance (SA). Owing to its large scale and diverse content, LasHeR facilitates the most comprehensive evaluations among existing RGB-T tracking benchmarks.

VTUAV. The VTUAV [23] dataset is a large-scale, high-quality RGB-T tracking benchmark collected from UAV platforms. It contains 500 video sequences, totaling approximately 1.7 million high-resolution frame pairs, making it the largest and highest-resolution RGB-T tracking dataset currently available. The dataset covers diverse target categories, including pedestrians, vehicles, and animals, as well as a wide range of scenes, such as roads, parks, and oceans. In addition, VTUAV provides precise frame-level segmentation mask annotations, supporting comprehensive evaluations for multiple tasks, including short-term tracking, long-term tracking, and pixel-level segmentation.

We summarize the aforementioned RGB-T tracking datasets and provide detailed scenario annotations for each tracking challenge, as shown in Table 2. Additionally, we compare the attribute challenge coverage across the datasets, as presented in Table 3.

3. More Comparisons with State-of-the-art Methods

In this section, we provide additional performance comparisons between our method and state-of-the-art trackers, including attribute-based analysis and qualitative experiments.

3.1. Analysis Across Various Attributes

We evaluate the performance of MOSSTrack on the GTOT dataset with respect to various tracking challenge attributes. Since GTOT is relatively limited in scale, recent methods are infrequently benchmarked on this dataset. To ensure a fair comparison, we locally deploy four state-of-the-art open-source multimodal trackers—CAFormer [21], SUTrack [2], BAT [1], and CAT++ [16]. Figure 1 presents the success and precision plots for each attribute subset as well as for the entire GTOT dataset. The results show that

Table 3. Comparison of tracking challenge attributes across existing RGB-T tracking datasets.

Datasets	Tracking challenge attributes																				
	NO	PO	TO	HO	OV	LI	HI	AIV	LR	DEF	BC	SA	TC	MB	CM	FL	FM	SV	SO	TVS	ARC
GTOT	-	✓	✓	-	-	✓	-	-	-	✓	-	-	✓	-	-	-	✓	✓	✓	-	-
RGBT210	✓	✓	✓	-	-	✓	-	-	✓	✓	✓	-	✓	✓	✓	-	✓	✓	-	-	-
RGBT234	✓	✓	✓	-	-	✓	-	-	✓	✓	✓	-	✓	✓	✓	-	✓	✓	-	-	-
LasHeR	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	-	-	✓
VTUAV	-	✓	✓	-	✓	✓	✓	-	✓	✓	✓	-	✓	✓	✓	-	✓	✓	-	✓	-

Table 4. Challenge -based PR/SR(%) analysis of MOSSTrack and 9 state-of-the-art algorithms on RGBT234. The top two results are highlighted in red and blue, respectively.

	VIPT [25]	TBSI [8]	CAT++ [16]	SDSTrack [5]	UnTrack [17]	XTrack [20]	CAFormer [21]	STTrack [6]	TBSI-Ext [9]	MOSSTrack (Ours)
NO	91.3/68.7	96.1/72.8	94.9/67.6	93.8/72.4	94.8/74.7	96.1/74.4	96.9/76.6	**98.6/76.9**	96.8/73.9	**99.4/78.4**
PO	85.4/63.0	88.7/64.7	88.9/62.7	82.9/63.0	83.7/63.2	87.9/66.3	88.6/67.9	89.0/67.3	**91.5/67.3**	**94.8/73.0**
TO	77.7/56.2	81.5/58.6	74.4/52.2	77.9/57.2	76.9/57.0	80.0/59.1	81.2/61.0	81.9/61.0	**87.9/63.7**	**86.2/66.2**
LI	81.0/58.4	89.2/63.6	80.9/55.5	85.7/63.5	78.2/55.8	87.3/63.6	88.7/**66.1**	89.1/64.4	**91.5**/65.8	**95.7/71.8**
LR	83.0/59.3	85.1/60.0	86.0/58.6	83.3/60.9	83.1/59.2	87.2/62.8	86.2/63.7	80.1/59.0	**90.7/64.6**	**87.6/67.6**
TC	83.0/62.2	85.8/63.2	85.0/61.5	83.7/62.0	89.1/67.4	89.9/67.0	89.8/68.1	**95.2/71.6**	91.0/68.2	**95.9/72.2**
DEF	81.7/62.2	84.1/63.7	80.4/56.7	80.4/62.2	81.2/63.3	83.2/64.2	83.6/65.6	83.0/64.6	**88.2/66.3**	**87.4/69.1**
FM	80.2/58.8	81.4/58.7	83.3/54.9	82.8/59.9	86.8/63.5	82.3/62.5	86.3/**64.5**	82.5/62.2	**88.4**/63.7	**87.3/64.9**
SV	83.8/63.0	89.9/66.8	84.8/59.3	84.7/64.3	82.9/63.3	85.5/64.9	87.5/67.7	91.3/69.4	**93.5/70.2**	**93.9/72.7**
MB	83.2/62.5	88.1/64.9	76.5/55.4	83.1/64.5	89.3/69.5	90.0/70.1	91.2/68.8	88.3/**70.4**	**91.6**/68.2	**92.3/74.1**
CM	83.0/62.0	88.0/65.0	77.9/56.2	83.4/63.0	87.0/66.3	89.0/67.6	89.4/67.2	87.8/**68.0**	**90.4**/67.2	**92.9/73.3**
BC	79.6/55.6	83.4/57.8	82.0/55.3	81.5/58.4	85.1/58.7	86.8/62.1	**87.0/62.9**	82.2/58.0	84.8/59.7	**87.6/65.1**
ALL	83.5/61.7	87.1/63.7	84.0/59.2	84.8/62.5	84.2/62.5	87.4/64.9	88.3/66.4	89.8/66.7	**91.0/67.0**	**94.3/71.5**

MOSSTrack consistently outperforms existing state-of-the-art methods in overall performance and across most challenge subsets.

In addition, we further evaluate MOSSTrack on the RGBT234 dataset under various attribute-specific challenges and compare it with nine multimodal trackers, including VIPT [25], TBSI [8], CAT++ [16], SDSTrack [5], UnTrack [17], XTrack [20], CAFormer [21], STTrack [6], and TBSI-Ext [9]. The results are presented in Table 4. MOSSTrack achieves particularly strong performance in the LI (Low Illumination) and TC (Thermal Crossover) scenarios, reaching PR/SR scores of 95.7%/71.8% and 95.9%/72.2%, respectively. In these two types of scenarios, key discriminative cues dynamically shift between the RGB and TIR modalities, placing higher demands on the tracker's ability to fully exploit cross-modal complementarity. MOSSTrack demonstrates stable performance under these conditions, highlighting its robustness in handling such challenges. Moreover, MOSSTrack maintains consistently high tracking accuracy in the DEF (Deformation) and SV (Scale Variation) scenarios, where the target undergoes substantial non-rigid shape deformation or significant changes in scale. Such conditions often challenge existing trackers, as large appearance variations can easily mislead feature extraction and lead to tracking drift. MOSSTrack achieves reliable target localization by effectively modeling temporal dependencies and adaptively refining spatial representations throughout the tracking process. These results further highlight the strong robustness and generalization capability of our spatio-temporal modeling strategy.

3.2. More Qualitative Results

In the supplementary material, we provide a more comprehensive qualitative comparison between MOSSTrack and three other advanced trackers—XTrack [20], UnTrack [17], and SDSTrack [5]—as shown in Fig. 2. To evaluate tracking performance, we selected four representative sequences from the LasHeR dataset, each representing distinct challenging scenarios.In the first sequence, a TIR frame dropout occurs at frame 281. Despite the missing thermal information, MOSSTrack remains stable and consistently maintains accurate target localization. In the second sequence, the target is immersed in a cluttered background with severe ambiguity, making it difficult to distinguish. Nevertheless, our approach exhibits remarkable stability and accuracy. In the third and fourth sequences, the target undergoes significant non-rigid deformation and is interfered with by objects of similar appearance. Leveraging its spatiotemporal modeling capability, MOSSTrack effectively adapts to these complex variations and preserves strong discriminative power. These qualitative results further highlight the superiority of our approach in handling diverse challenging tracking attributes.

Figure 2. Qualitative comparison between MOSSTrack and three other trackers on challenging tracking cases.

References

[1] Bing Cao, Junliang Guo, Pengfei Zhu, and Qinghua Hu. Bidirectional adapter for multimodal tracking. In *Proceedings of the AAAI conference on artificial intelligence*, pages 927–935, 2024. 3

[2] Xin Chen, Ben Kang, Wanting Geng, Jiawen Zhu, Yi Liu, Dong Wang, and Huchuan Lu. Sutrack: Towards simple and unified single object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2239–2247, 2025. 3

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 1

[4] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. Lasot: A high-quality benchmark for large-scale single object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5374–5383, 2019. 1

[5] Xiaojun Hou, Jiazheng Xing, Yijie Qian, Yaowei Guo, Shuo Xin, Junhao Chen, Kai Tang, Mengmeng Wang, Zhengkai Jiang, Liang Liu, et al. Sdstrack: Self-distillation symmetric adapter learning for multi-modal visual object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26551–26561, 2024. 1, 4

[6] Xiantao Hu, Ying Tai, Xu Zhao, Chen Zhao, Zhenyu Zhang, Jun Li, Bineng Zhong, and Jian Yang. Exploiting multimodal spatial-temporal patterns for video object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3581–3589, 2025. 1, 4

[7] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE transactions on pattern analysis and machine intelligence*, 43(5):1562–1577, 2019. 1

[8] Tianrui Hui, Zizheng Xun, Fengguang Peng, Junshi Huang, Xiaoming Wei, Xiaolin Wei, Jiao Dai, Jizhong Han, and Si Liu. Bridging search region interaction with template for rgb-t tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13630–13639, 2023. 4

[9] Bo Li, Fengguang Peng, Tianrui Hui, Xiaoming Wei, Xiaolin Wei, Lijun Zhang, Hang Shi, and Si Liu. Rgb-t tracking with template-bridged search interaction and target-preserved template updating. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(1):634–649, 2025. 4

[10] Chenglong Li, Hui Cheng, Shiyi Hu, Xiaobai Liu, Jin Tang, and Liang Lin. Learning collaborative sparse representation for grayscale-thermal tracking. *IEEE Transactions on Image Processing*, 25(12):5743–5756, 2016. 2

[11] Chenglong Li, Nan Zhao, Yijuan Lu, Chengli Zhu, and Jin Tang. Weighted sparse representation regularized graph learning for rgb-t object tracking. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1856–1864, 2017. 3

[12] Chenglong Li, Xinyan Liang, Yijuan Lu, Nan Zhao, and Jin Tang. Rgb-t object tracking: Benchmark and baseline. *Pattern Recognition*, 96:106977, 2019. 3

[13] Chenglong Li, Wanlin Xue, Yaqing Jia, Zhichen Qu, Bin Luo, Jin Tang, and Dengdi Sun. Lasher: A large-scale high-diversity benchmark for rgbt tracking. *IEEE Transactions on Image Processing*, 31:392–404, 2021. 1, 3

[14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1

[15] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 1

[16] Lei Liu, Chenglong Li, Yun Xiao, Rui Ruan, and Minghao Fan. Rgbt tracking via challenge-based appearance disentanglement and interaction. *IEEE Transactions on Image Processing*, 33:1753–1767, 2024. 3, 4

[17] Yinchao Ma, Yuyang Tang, Wenfei Yang, Tianzhu Zhang, Jinpeng Zhang, and Mengxue Kang. Unifying visual and vision-language tracking via contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4107–4116, 2024. 4

[18] Matthias Muller, Adel Bibi, Silvio Giancola, Salman Alsubaihi, and Bernard Ghanem. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *Proceedings of the European conference on computer vision (ECCV)*, pages 300–317, 2018. 1

[19] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019. 1

[20] Yuedong Tan, Zongwei Wu, Yuqian Fu, Zhuyun Zhou, Guolei Sun, Chao Ma, Danda Pani Paudel, Luc Van Gool, and Radu Timofte. Xtrack: Multimodal training boosts rgb-x video object trackers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 4

[21] Yun Xiao, Jiacong Zhao, Andong Lu, Chenglong Li, Bing Yin, Yin Lin, and Cong Liu. Cross-modulated attention transformer for rgbt tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8682–8690, 2025. 3, 4

[22] Botao Ye, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Joint feature learning and relation modeling for tracking: A one-stream framework. In *European conference on computer vision*, pages 341–357. Springer, 2022. 1

[23] Pengyu Zhang, Jie Zhao, Dong Wang, Huchuan Lu, and Xiang Ruan. Visible-thermal uav tracking: A large-scale benchmark and new baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8886–8895, 2022. 3

[24] Xiaosong Zhang, Yunjie Tian, Lingxi Xie, Wei Huang, Qi Dai, Qixiang Ye, and Qi Tian. Hivit: A simpler and more efficient design of hierarchical vision transformer. In *The eleventh international conference on learning representations*, 2023. 1

[25] Jiawen Zhu, Simiao Lai, Xin Chen, Dong Wang, and Huchuan Lu. Visual prompt multi-modal tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9516–9526, 2023. 4